Cookies!

Benjamin Agyei, Christian Miller, Elliot Shannon, Grayson White

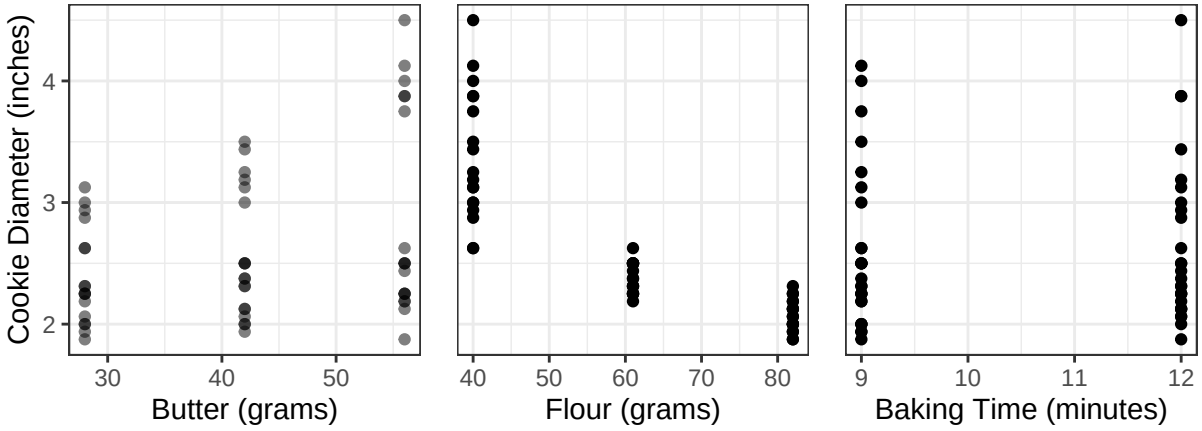
2022-12-07

Introduction

Methods

Results

Once we had baked and measured all of our cookies, it was time to begin modeling with our data. Before diving deep into modeling, we first took the important step of exploratory data analysis. To get a good sense of our data, we plotted each explanatory variable with the response variable. Figure X shows these relationships.



Initially, we see that the relationship between cookie diameter and butter is moderate in strength and positive. The linearity of this relationship needs to be examined further, but generally the trend is somewhat linear. Further, the relationship between cookie diameter and flour seems to be even stronger, but now a negative relationship. This tells us that as we increase the flour in a cookie, we would expect the cookie diameter to decrease (and of course, the converse for butter). The relationship between cookie diameter and time in oven is extremely hard to discern, and from Figure X we expect time in oven to have a very minimal effect on cookie diameter, if any.

Now, we move to modeling. We first fit the candidate model described in the Introduction, Model 1,

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \epsilon_i$$

where x_{1i} corresponds to the ith observation of flour, x_{2i} the ith observation of butter, and x_{3i} the ith observation of time in oven. We fit the model in R and assess Model

1 via the quartet of diagnostic plots provided by the `gglm` R package, and by metrics from the model summary. Figure XX displays Model 1's diagnostic plot quartet.

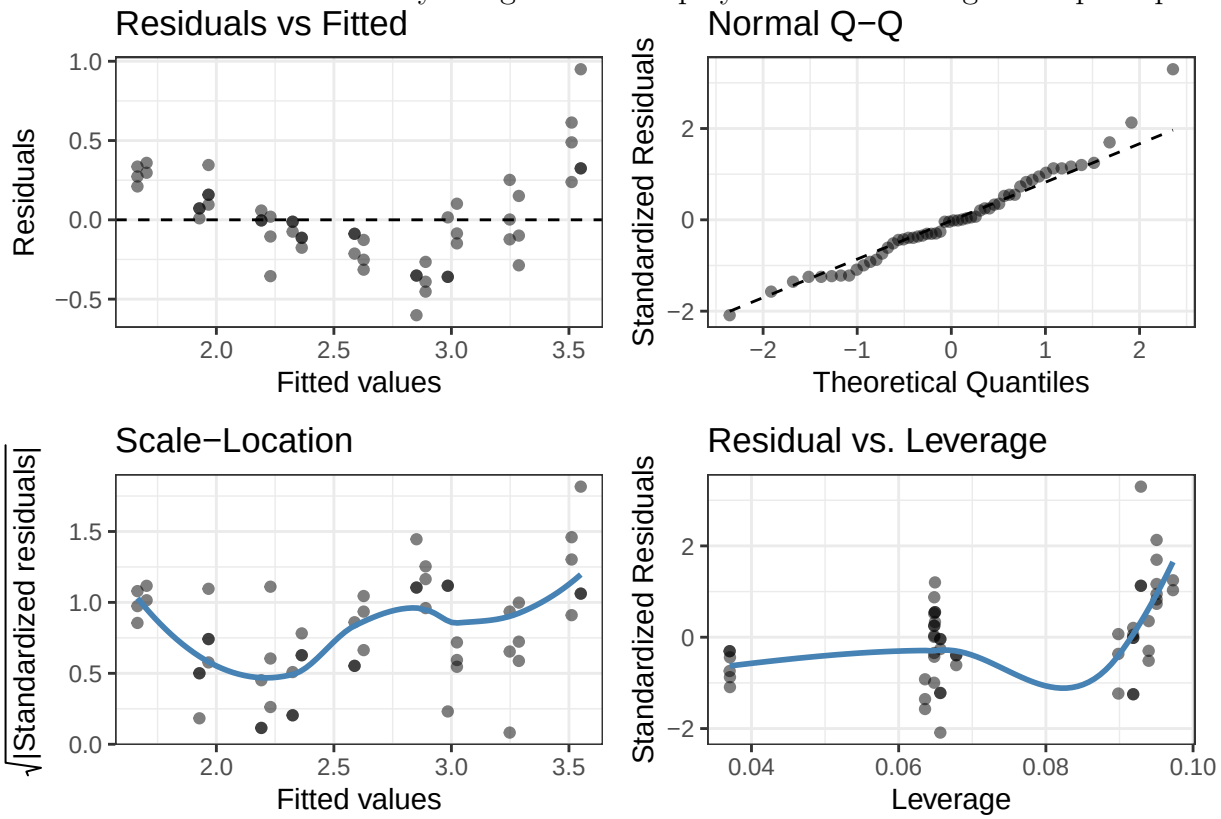


Figure XX shows some departures from essential modeling conditions. Most notably, we see a U-shaped pattern in the residuals vs fitted plot. This violates the assumption that the error terms are independent of each other, and tells us that there is some curvature in some relationship between the response and an auxiliary variable that we are not accounting for in Model 1. Further, we see some higher leverage points with large residuals. These points could be highly influential in determining the model parameter estimates, and may be causing some bias in our model parameter estimates. Outside of these issues, the other model diagnostics look good for the initial candidate model. We also find that Model 1's model fit is quite good to begin with, with an R^2 value of 0.802. We find that for Model 1, all variables expect time in oven are significant at the 0.001 level, and time in oven is not significant at all.

We now move to attempting to fix the issues made apparent by the diagnostic plots, while maintaining good model fit. Largely, we had two approaches to addressing the diagnostic issues. Both approaches remove the time in oven variable as it was seen to not benefit the model, while adding the complexity of another variable. The first approach considers adding both a multiplicative interaction term between butter and flour, and a squared term for butter. We fit Model 2 as follows:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \epsilon_i$$

where xji is the same as Model 1 for $j \in \{1, 2\}$, x_{3i} corresponds to the ith observation of butter times flour, and x_{4i} corresponds to the ith observation of the square of butter.

Again, we fit Model 2 in R and initially look at the diagnostic plots shown in Figure XXX.

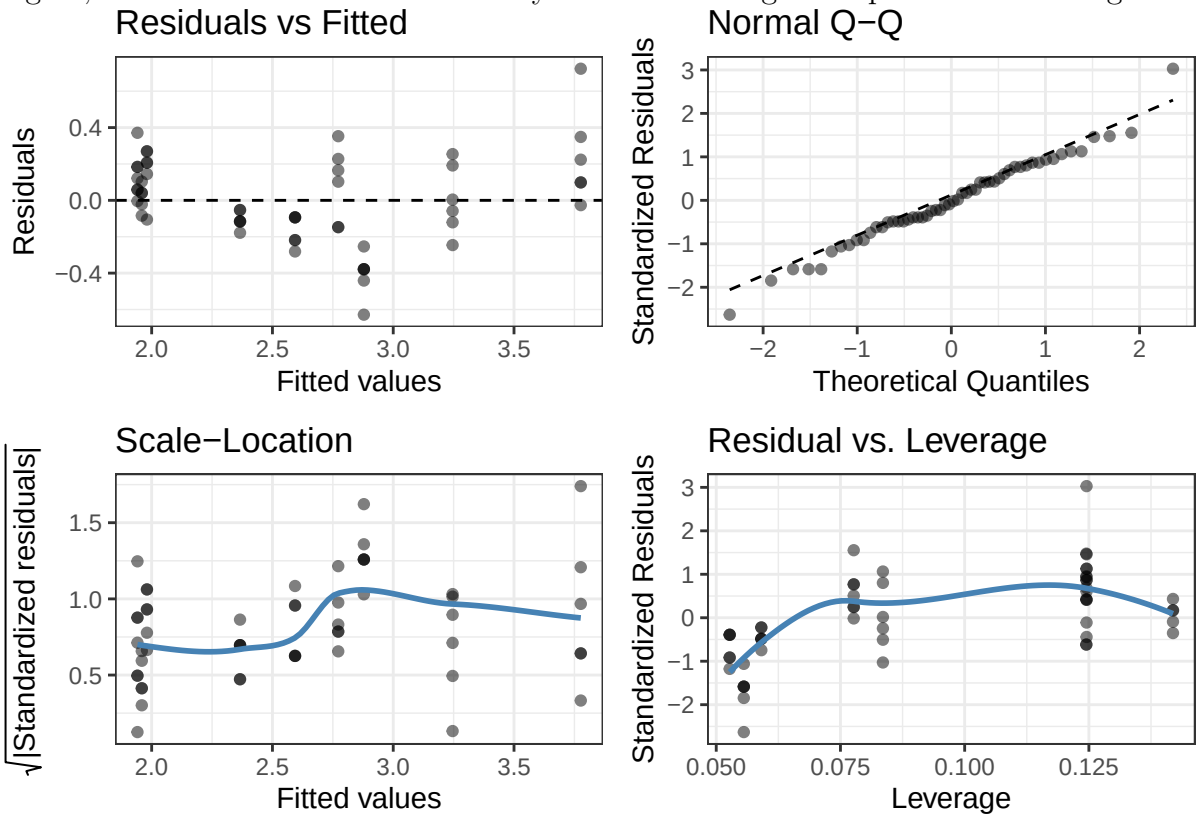
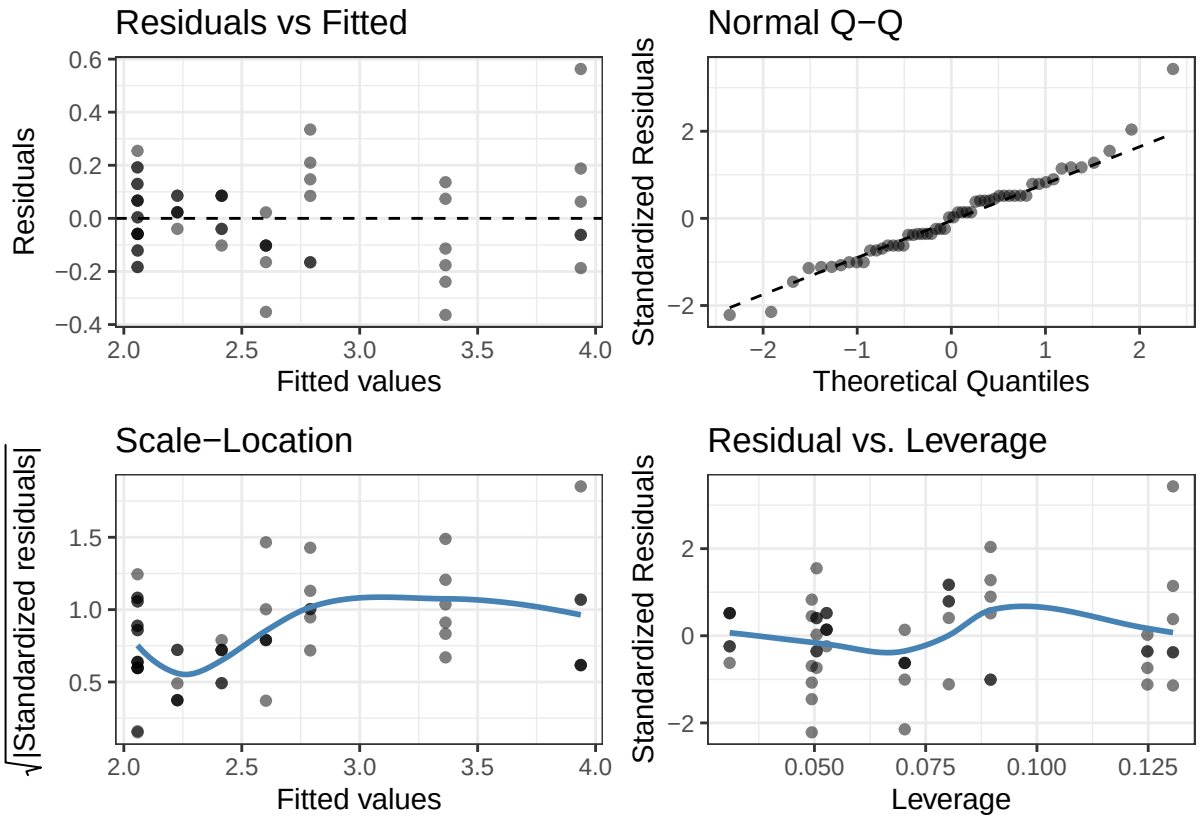


Figure XXX shows that Model 2 begins to address the issue of the U-shaped residuals and high leverage-high residual points, however the residuals seem to still follow a vaguely U-shaped pattern. With Model 2, we find that the only term significant at the 0.05 level is the flour times butter term as it seems to soak up similar attributes in the model to butter and flour alone, along with the square of butter. This tells us there might be an issue with multicollinearity, which could cause issues with model interpretability and the stability of the estimates for our β's. The R^2 value for Model 2 was 0.864, which is an increase in R^2 compared to Model 1. We also see an adjusted R^2 increase from 0.790 to 0.853 when comparing with Model 1, leading us to believe that adding these terms is still improving model fit.

Our final model, Model 3, takes a different approach at correcting model diagnostics and improving fit, all while maintaining a high level of interpretability. For Model 3, we first fit the model:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \epsilon_i$$

Model 3 is fit with only three auxiliary variables, where x_{1i} corresponds to the ith observation of butter, x_{2i} corresponds to the ith observation of flour, and x_{3i} corresponds to ith observation of the ratio of butter to flour. We believe that the ratio of butter to flour explains the change in cookie diameter with high interpretability, as the amount of butter and amount of flour are thought to explain the change in cookie diameter, but how they effect cookie diameter likely changes based on how much of the other is already in the cookie. Plotting the ratio of flour to butter against cookie diameter shows a much stronger relationship than their product, and so we fit Model 3 with this ratio. Figure XXXX show the model diagnostics for Model 3.



We see that any discernible pattern in the residual vs fitted plot is gone, and all other diagnostic plots look satisfactory. All variables included in Model 3 were significant at the 0.05 level, and the R^2 value for Model 3 is 0.937 with an adjusted R^2 of 0.932. These metrics tell us that the model fit is quite good.

Discussion

Conclusion

References